SemSEDoc: Utilización de tecnologías semánticas en el aprovechamiento de los repositorios documentales de los proyectos de desarrollo de software

Francisco J. García-Peñalvo
Departamento de Informática y Automática, Universidad de Salamanca, Salamanca, España
Ricardo Colomo-Palacios
Departamento de Informática, Universidad Carlos III de Madrid, Madrid, España
Pedro Soto-Acosta
Departamento de Organización de Empresas y Finanzas, Universidad de Murcia, Murcia, España
Isabel Martínez-Conesa
Departamento de Economía Financiera y Contabilidad, Universidad de Murcia, Murcia, España
Enric Serradell-López
Departamento de Estudios en PIMEC-SEFES, Universidad Oberta de Catalunya, Catalunya, España

Introducción. *Debido a la importancia de los sistemas de información en la sociedad actual, el desarrollo de los mismos supone una actividad clave para la sociedad del conocimiento. En este entorno, los sistemas son cada vez más complicados y su desarrollo implica la generación de una documentación de soporte a este proceso que debe ser gestionada de forma efectiva y eficiente.*
Método. *A partir de la necesidad de contar con repositorios documentales de soporte al desarrollo de software, se propone la utilización de las tecnologías semánticas para la catalogación, etiquetado, búsqueda y visualización de los artefactos software almacenados en los repositorios.*
Resultados. *Los resultados de su implantación se*

consideran altamente satisfactorios. Por una parte, el feedback recibido sobre las técnicas empleadas para el anotado y la visualización de los resultados es positivo y, en segundo lugar, las pruebas realizadas (Precision, Recall y F1) resultan muy alentadoras para la obtención de resultados fiables.
Conclusiones. *Las tecnologías semánticas por su creciente grado de madurez proporcionan un marco de explotación eficiente y eficaz para los repositorios documentales generados en los proyectos de desarrollo de software.*

Abstract in English

change font

Introduction

La tecnología se ha convertido en un agente de primer orden en la Sociedad del Conocimiento debido a su importancia en el entorno de las organizaciones y para el funcionamiento de la sociedad en general (Soto-Acosta y Meroño-Cerdan 2008, Soto-Acosta *et al.* 2010b). En las organizaciones actuales, que cada vez son más intensivas en conocimiento, la tecnología representa la columna vertebral que hace que el conocimiento y la información se puedan compartir, crear y almacenar, posibilitando a las organizaciones la oportunidad de ventajas competitivas basadas en el conocimiento (Lopez-Nicolas and Soto-Acosta 2010; Sharma *et al.* 2010). Estos hechos han colocado al conocimiento como un activo clave en torno al cual se han articulado términos y conceptos tales como: "Sociedad del Conocimiento", "Economía del Conocimiento" o "Cultura del Conocimiento" (Bakry y Alfantookh 2010).

En este entorno, las tecnologías de la información y la comunicación conforman

3 un activo esencial a través del cual se realizan actividades intensas en conocimiento en las que las organizaciones cimentan sus operaciones (Trigo *et al.* 2009). En este sentido, las tecnologías de la información y la comunicación son consideradas fundamentales para el desarrollo de productos y servicios (Gonzálvez-Gallego *et al.* 2010; Soto-Acosta *et al.* 2010c). El sector de las tecnologías de la información y la comunicación lo componen organizaciones que se dedican a: desarrollar y comercializar software empaquetado y a medida; a ofrecer infraestructura tecnológica; y a dar soporte tecnológico y consultoría (O'Sullivan and Dooley 2010). La importancia que tiene el software en el ámbito de las tecnologías de la información y la comunicación ha convertido a la industria de desarrollo de software en una de las más influyentes en el mundo y en una industria clave para el crecimiento económico. El proceso del desarrollo de software se basa en tres pilares fundamentales: procesos, tecnologías y personas. Estos pilares se encuentran interconectados, formando un triángulo fundamental para que las organizaciones operen (Hernández-Lopez *et al.* 2010; Soto-Acosta *et al.* 2010a). Sin embargo, el desarrollo de software implica, además de la producción de programas ejecutables y otros elementos de software, la generación de documentación de soporte al proceso de desarrollo, a la explotación y al mantenimiento del software. En muchas ocasiones, la documentación de soporte no presenta la calidad que sí se encuentra en el producto software final (Lethbridge *et al.* 2003). Sin embargo éste elemento es una herramienta útil en dos tareas fundamentales del proceso de desarrollo de software: el mantenimiento (e.g., Kajko-Mattsson 2005) y la reutilización (e.g., Yao *et al.* 2008).

La documentación es un componente fundamental para la calidad del software (Kajko-Mattsson 2005), facilitando el mantenimiento perfectivo, defectivo y adaptativo. La literatura ha demostrado desde los años 1980 que la ausencia de documentación es una de las causas principales de defectos en el mantenimiento, al mismo tiempo que los costes de mantenimiento de los sistemas que cuentan con una buena documentación son menores (e.g., Card *et al.* 1987; Rombach y Basili 1987). Además, una buena documentación permite habilitar la reutilización del software. La reutilización del software es el proceso de crear software desde sistemas existentes en lugar de realizarlo desde cero (Krueger 1992). La reutilización es una técnica de probada efectividad para el ahorro de costes, ya sea, entre otros aspectos, en los tiempos de desarrollo o en la fiabilidad de los componentes previamente probados (Frakes y Kang 2005; Kim y Stohr 1998; Mohagheghi y Conradi 2007; Selby 2005).

En el entorno de la reutilización, los mecanismos de soporte a la búsqueda con vistas a la reutilización de software suelen estar basados en palabras clave (Sugumaran y Storey 2003). Sin embargo, atendiendo a Yao *et al.* (2008), estos mecanismos pueden ser enriquecidos de forma sustancial con el uso de técnicas basadas en el conocimiento que soporten una mejor búsqueda en los repositorios software.

Con el propósito de que tanto la reutilización como el mantenimiento del software se puedan realizar de forma más efectiva, el presente trabajo presenta

4 la herramienta SemSEDoc (Semantic Software Engineering Documentation). Esta herramienta propone la utilización de tecnologías semánticas en el aprovechamiento de los repositorios documentales de los proyectos de desarrollo de software. Para ello, SemSEDoc mediante el uso de Procesamiento del Lenguaje Natural, lleva a cabo un análisis de la documentación generada, anotando semánticamente los documentos gracias al uso de una ontología de dominio y otra funcional.

Estado del arte

El advenimiento de la Web semántica ha supuesto una revolución en la forma de acceso y almacenamiento de la información. El término Web Semántica fue acuñado por Berners-Lee *et al.* (2001) para describir la evolución desde un paradigma basado en documentos hacia un nuevo paradigma que incluye de forma coordinada la información y los datos con el propósito de ser manipulado de forma automática. La utilización de soporte semántico en soluciones de las tecnologías de la información y la comunicación permite la introducción de inteligencia en los sistemas software, circunstancia que no es posible en los sistemas basados en datos simples (Álvarez-Sabucedo 2010). En particular, atendiendo a Berners-Lee *et al.* (2006), las tecnologías semánticas se entienden para la creación de codificaciones declarativas con el fin de facilitar la interoperatividad, la integración y el acceso a los datos. Teniendo en cuenta que el acceso a la información es uno de los retos más importantes para los sistemas de información en la actualidad (Morales-Del Castillo *et al.* 2009), las tecnologías semánticas se perfilan como habilitadoras del acceso inteligente y preciso a grandes repositorios de datos.

Atendiendo a Warren (2006), las tecnologías semánticas proporcionan una visión complementaria que, en muchos casos, ha expandido y reemplazado los arquetipos de la gestión del conocimiento y de la información (Davies *et al.* 2007). Las ontologías proporcionan vocabularios estructurados que describen especificaciones formales de las conceptualizaciones. Estas tecnologías proponen una solución para representar el significado de la información, lo que conduce a una más efectiva gestión de los datos gracias al establecimiento de un vocabulario común (Shadbolt *et al.* 2006). Los beneficios que conllevan la adición de semántica a los contenidos consisten en eliminar las inconsistencias terminológicas. La anotación semántica identifica formalmente conceptos y relaciones entre éstos (Uren *et al.* 2006). Esta anotación debe ser explícita, formal y no ambigua, siendo sus beneficios principales una mejor búsqueda e interoperabilidad (Uren *et al.* 2006). Una introducción sobre las principales herramientas y conceptos relativos al empleo de tecnologías semánticas se puede encontrar en el trabajo de Brooks (2009).

El empleo de las tecnologías semánticas en aplicaciones industriales se ha extendido, por su utilidad e impacto (García-Crespo *et al.* 2010a). Así, su aplicación en el campo de la ingeniería del software no es nueva. La literatura ha recogido esfuerzos relativos a la aplicación de tecnologías semánticas en campos como: la formación y gestión de equipos de desarrollo de software (e.g., Colomo-

5 Palacios *et al.* 2010; Valencia-García *et al.* 2010); la utilización de patrones (Dietrich *et al.* 2008); componentes (Colomo-Palacios *et al.* 2008) o como tecnología habilitadora para la reutilización (Henriksson *et al.* 2008); la gestión de métricas del software (e.g., Gall *et al.* 2008; García-Crespo *et al.*, 2009); el soporte a los procesos de análisis (Girardi y Leite 2008; Tappolet *et al.* 2010); apoyo en el desarrollo de software global (Wongthongtham *et al.* 2009); la ayuda para la evaluación CMMi (Lee and Wang 2009), por citar algunos de los casos y áreas más significativos y recientes. El trabajo de Zhao *et al.* (2009) contiene una revisión detallada de la aplicación de técnicas semánticas en diferentes aspectos relacionados con la ingeniería del software.

En el entorno específico del enriquecimiento semántico de la documentación generada durante el proceso de desarrollo de software, los esfuerzos también son notables y recientes (e.g., De Lucia *et al.* 2007; García *et al.* 2009; Hyland-Wood *et al.* 2008; Zhang *et al.* 2008). Sin embargo, ninguno de los trabajos anteriores posibilita el etiquetado semántico semi-automático de la documentación generada como soporte al proceso software, mediante una ontología funcional y otra de dominio. Dicha funcionalidad permite una búsqueda y gestión de la documentación más efectiva, habilitando de esta manera un mejor mantenimiento de la propia documentación y del software soportado por la misma.

SemSEDoc: arquitectura y principales funcionalidades

SemSEDoc se presenta como una herramienta de apoyo a la gestión de la documentación almacenada en repositorios de proyectos. La herramienta ha sido diseñada para actuar en proyectos realizados bajo la metodología Métrica versión 3. Métrica3 es una metodología de desarrollo elaborada por el Consejo Superior de Informática del Ministerio de Administraciones Públicas de España. La metodología presenta tres procesos principales: A) Planificación, B) Desarrollo y C) Mantenimiento de Sistemas de Información. De estos tres, se ha considerado únicamente el proceso de Desarrollo de Sistemas de Información. Dentro de este proceso se han tenido en cuenta la documentación que surge de los cinco subprocesos que lo componen: A) Estudio de Viabilidad del Sistema; B) Análisis del Sistema de Información; C) Diseño del Sistema de Información; D) Construcción del Sistema de Información e Implantación; y E) Aceptación del Sistema. En la figura 1, se plasma la estructura de la herramienta y su relación con el repositorio de proyectos y los usuarios.

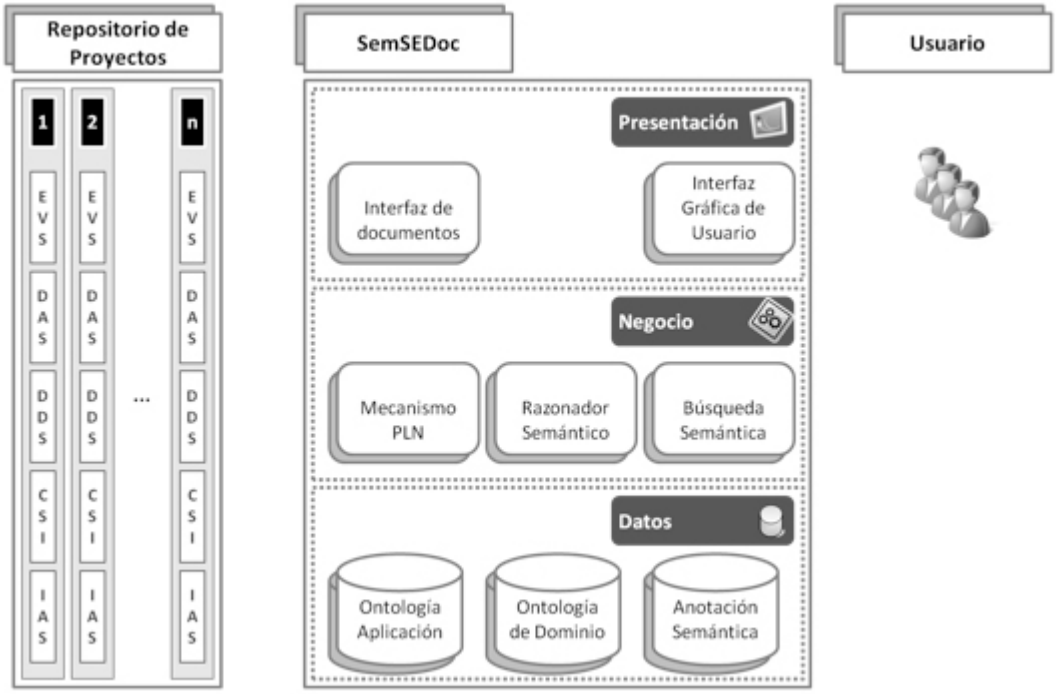

Figura 1: Arquitectura de SemSEDoc

SemSEDoc obtiene la información necesaria del proyecto mediante una interfaz de documentos con el repositorio de proyectos. Dicho repositorio contiene la documentación generada en diversos proyectos a partir de las recomendaciones de Métrica3. La documentación es analizada mediante un Mecanismo de PLN que detecta aquellos elementos susceptibles de ser etiquetados, proceso que se lleva a cabo de forma semi-automática. Teniendo en cuenta que la anotación semántica, debido al esfuerzo que conlleva, se considera una barrera potencial para el uso de tecnologías semánticas (Benjamins *et al.* 2008), se ha optado por una anotación semi-automática, con el fin de hacer menos tedioso el proceso y evitar los errores de las anotaciones semánticas automáticas.

Una vez llevada a cabo la etiquetación semántica, que se almacena debidamente en la capa correspondiente, los mecanismos de búsqueda y razonamiento semántico se encuentran habilitados para llevar a cabo su labor, directamente relacionada con las capacidades de la herramienta de buscar elementos relativos a proyectos de desarrollo de software y su interrelación con otros. Recordemos que el fin último es el de incrementar la reutilización y de acrecentar el mantenimiento del software.

La arquitectura del aplicativo se compone de tres capas: Presentación, Negocio y Datos. Esta solución arquitectónica garantiza un mejor mantenimiento, además de permitir un desarrollo más modular. La capa de Presentación se comunica dentro de la aplicación con la capa de Negocio y sirve de interfaz de SemSeDoc con dos actores cruciales del sistema. En primer lugar, se comunica con los usuarios a través del módulo “Interfaz Gráfica de Usuario”. Este módulo aglutina la interacción con el usuario final en relación a las búsquedas y la anotación semántica de documentos. El segundo de los módulos es el encargado de interactuar con el repositorio de proyectos. En la actualidad, dicho módulo es una interfaz que lleva a cabo la comunicación de la aplicación con el conjunto de

7 documentos listados en un cúmulo de proyectos. El requisito para su utilización es que dichos documentos deben encontrarse en un índice y su contenido debe ser codificado en formato HTML. Esta funcionalidad se plantea ser sustituida por la implementación de un crawler que permita una búsqueda no estructurada de documentos codificados en diferentes formatos (Portable Document Format, Rich Text Format, Microsoft Word, etc).

En la capa de lógica de negocio existen tres módulos. El primero de ellos es el correspondiente al mecanismo para el PLN. Este componente analiza los documentos que se encuentran en el repositorio de proyectos. Así, el componente utiliza las herramientas GATE para la anotación sintáctica de los contenidos y JAPE para extraer los conceptos relativos a la ontología de dominio y la funcional. Posteriormente, a partir de la lista proporcionada se sugiere la anotación semántica de los diferentes elementos de forma semi-automática. El segundo de los módulos que presenta la capa de lógica de negocio es el razonamiento semántico. Un razonador semántico típicamente deriva hechos de una base de conocimiento con el propósito último de formular nuevas conclusiones. En el caso particular de la herramienta que nos ocupa consiste en un razonador basado en OWL (Ontology Web Language). Este lenguaje incorpora un modelo semántico que permite la creación de sistemas de razonamiento. Así, la herramienta incorpora a RACER (Renamed ABox and Concept Expression Reasoner) para llevar a cabo la tarea descrita. El componente de búsqueda semántica utiliza el lenguaje de consultas SPARQL RDF para llevar a cabo consultas en la capa de datos. Una de las características de la herramienta es la búsqueda por facetas. Con metadatos por facetas (Ranganathan 1962), el espacio de información se divide en particiones utilizando dimensiones conceptuales ortogonales de los datos. Estas dimensiones se llaman facetas y representan las características de los elementos de la información. Estas facetas se utilizan para seleccionar o filtrar elementos relevantes en un determinado espacio de información, llevando así a los usuarios exactamente a la información necesaria. Estas facetas son las propiedades definidas en el dominio de las ontologías. La búsqueda semántica por facetas ha sido utilizada con éxito en herramientas del ámbito semántico, tal y como reflejan diferentes trabajos en la literatura (e.g., García-Crespo *et al.* 2010b, 2011; Gómez-Berbís *et al.* 2011; Prasad y Madalli 2008; Suominen *et al.* 2009).

La última capa es la de datos o persistencia. En dicha capa se almacenan, por un lado, las ontologías de aplicación y de negocio y, por otro, las anotaciones semánticas de los documentos analizados. Dicho repositorio ha sido implementado utilizando las capacidades de Jena sobre SESAME. De esta forma, se consigue la persistencia de las ontologías, la realización de consultas en el entorno desarrollado y se ofrece un nivel de abstracción de la complejidad que permite el almacenamiento y la recuperación de ontologías OWL DL en conjunción con su sintaxis RDF.

A continuación se describe la evaluación llevada a cabo con el propósito de testar si SemSEDoc sirve como sistema de gestión de la documentación

8 generada como soporte al proceso de desarrollo de software en un entorno controlado.

Diseño

Una vez que el sistema ha sido desarrollado y testado desde el punto de vista del proceso de desarrollo, se necesita probar la validez de SemSEDoc. Así, se pretende conocer dos tipos de características del aplicativo en relación con su utilización. En primer lugar, la bondad de los resultados del proceso de búsqueda y, por ende, del conjunto de funcionalidades de etiquetado y, en segundo lugar, se pretende testar si el procedimiento de anotación semiautomático produce sugerencias de anotación pertinentes de acuerdo con el juicio de los expertos. Esta segunda característica comprueba si las sugerencias de anotación de los diferentes documentos coincide con la anotación manual de los expertos.

Ambos aspectos requieren la utilización de métricas conocidas y efectivas. Para ello, se utilizarán las métricas Precision, Recall y F1. Las dos primeras fueron introducidas por Cleverdon *et al.* (1966). Con posterioridad, la métrica F1 fue diseñada por van Rijsbergen (1979) con el propósito de integrar en una única medida y con importancia equivalente las capacidades de ambas métricas. Un trabajo reciente y relevante que analiza el uso de las citadas métricas es el publicado por Good and Tennis (2009). Las tres métricas se pueden definir atendiendo a las siguientes formulaciones:

Precision = Resultados encontrados correctos / Total de resultados encontrados

Recall = Resultados encontrados correctos / Total de resultados Correctos

F1 = (2 * Precision*Recall) / (Precision+Recall)

Para llevar a cabo ambos procedimientos se contó con un repositorio de proyectos compuesto por un total de 9 proyectos de desarrollo de software ejecutados bajo los dictados de Métrica3. Dicho repositorio de proyectos fue alojado en un ordenador conectado a una red de área local en la que se encontraba también el servidor que albergaba a SemSEDoc. Con el propósito de testar los aspectos expuestos anteriormente, se contó con un total de 18 sujetos cuya participación efectiva en los proyectos les habilitara permite realizar un juicio sobre la bondad de la herramienta en la búsqueda y anotación semántica.

Teniendo en cuenta el doble objetivo de la evaluación, se diseñó una tarea doble para los sujetos. Así, y tras llevar a cabo la recolección de los datos de identificación del sujeto, en primer lugar, cada sujeto debía llevar a cabo la anotación de uno de los documentos generados como soporte al desarrollo del proyecto en el que los sujetos habían llevado a cabo su labor. El documento (un acta de reunión relativa a la recolección de requisitos), que en todos los casos había sido examinado con anterioridad, contenía texto para llevar a cabo un conjunto suficiente de anotaciones (alrededor de veinte). El diseño del experimento habilitaba la verificación de la bondad del proceso de anotación en

9 relación al conjunto de anotaciones posibles mediante la comparación de las anotaciones de los expertos con las sugerencias producidas por el sistema. En segundo lugar, se solicitaba a los sujetos que llevaran a cabo una búsqueda de un término (un aspecto específico relativo a un requisito de usuario) restringiendo la misma a un conjunto de documentos dado (estudio de viabilidad del sistema y análisis del sistema de información). En este caso, el objetivo es probar si se producen los resultados de búsqueda deseados. Los datos de los resultados de búsqueda se comparan con la búsqueda realizada, por los sujetos de forma libre, mediante el acceso a los documentos utilizando un navegador y capacidades de búsqueda textual, además de la lectura.

Los sujetos llevaron a cabo su labor asistidos en todo momento por un miembro del equipo de desarrollo de SemSEDoc y codificaron sus resultados con ayuda de un cuestionario. A la finalización de su trabajo entregaron el cuestionario al miembro del equipo de proyecto, quien codificó los resultados utilizando la herramienta estadística SPSS.

Muestra

La muestra se compone de 18 sujetos. La distribución de los sujetos por sexos se establece en 13 hombres (72.22%) y 5 mujeres (27.78%). Todos ellos cuentan con experiencia en los 9 proyectos objeto de estudio, escogiendo 2 sujetos por proyecto, con el fin de contar con más de una opinión por proyecto. El rol de los diferentes sujetos en los proyectos se distribuye de la siguiente manera: Jefe de Proyecto (7), Gestor de Configuración (5), Gestor de Calidad (4) y Analista-Programador (2). Todos los sujetos contaban con experiencia previa en la anotación semántica de textos.

Resultados

La tabla 1 contiene los resultados de las anotaciones de los sujetos y la herramienta. La columna Sujeto x contiene las anotaciones realizadas por uno de los sujetos en un determinado proyecto. La columna TOTAL refleja la suma de las anotaciones diferentes de los dos sujetos que figuran en las columnas anteriores. SemSEDoc contiene las anotaciones de la herramienta y Comunes exhibe la intersección de las anotaciones de la herramienta y de la columna TOTAL.

Proyecto	Sujeto 1	Sujeto 2	TOTAL	SemSEDoc	Comunes
Proyecto 1	21	21	21	19	19
Proyecto 2	24	24	24	25	20
Proyecto 3	19	20	20	21	18
Proyecto 4	26	27	27	29	25
Proyecto 5	22	22	22	23	19

Proyecto 6	19	18	19	18	17
Proyecto 7	18	19	19	19	18
Proyecto 8	20	20	20	18	17
Proyecto 9	25	25	25	23	21

Tabla 1: Comparativa de las anotaciones llevadas a cabo por los sujetos y la herramienta.

A partir de los resultados mostrados en la Tabla 1, se localizó por parte de los sujetos un total de 197 elementos susceptibles de ser anotados. La herramienta detectó un total de 195 elementos, de los que 174 coincidían con los señalados por los usuarios. Con el propósito de verificar si existían diferencias significativas desde el punto de vista estadístico entre los resultados obtenidos por los expertos (TOTAL) y las anotaciones sugeridas por SemSEDoc que se consideran correctas (Comunes), se utilizó el método estadístico de la t de Student. El nivel de significación se situó en 0.05. Los resultados del test indican que ambos resultados no presentan diferencias significativas ($t(9)=2.02$, $p>0.05$) entre sí, por lo que se puede confirmar la cercanía de los resultados ofrecidos por SemSEDoc al estándar de anotación.

Con respecto a las métricas de bondad, la medida Precision (Precisión) se situó en 0.892307692, mientras que la métrica Recall (Exhaustividad) resultó ser de 0,883248731 y la ponderada armónica F1 fue de 0.887755102.

A continuación, se exponen los resultados de la comparativa entre las búsquedas llevadas a cabo por la herramienta desarrollada y la que han llevado a cabo los usuarios. La Tabla 2 contiene los resultados de dicho proceso. La descripción de las columnas es análoga a la que aparece para la Tabla 1, con la salvedad de que los resultados son consignados para las búsquedas en lugar de las anotaciones.

Proyecto	Sujeto 1	Sujeto 2	TOTAL	SemSEDoc	Comunes
Proyecto 1	45	50	52	48	41
Proyecto 2	66	70	71	72	62
Proyecto 3	30	23	31	34	18
Proyecto 4	28	23	28	29	19
Proyecto 5	52	47	54	43	43
Proyecto 6	58	69	75	54	50
Proyecto 7	43	36	44	42	35
Proyecto	32	34	34	29	24

8					
Proyecto 9	56	49	59	52	40

Tabla 2: Comparativa de las búsquedas llevadas a cabo por los sujetos y la herramienta.

De los resultados ofrecidos en la Tabla 2, se desprende que se localizan por parte de los sujetos un total de 448 resultados de búsqueda. La herramienta arrojó un total de 403 resultados, de los que 332 coinciden con los señalados por los usuarios. De forma análoga a la anterior, los resultados de la comparación de medias entre TOTAL y Comunes tampoco presenta diferencias estadísticamente significativas entre ambas poblaciones ($t(9)=1.73$, $p>0.05$). Este hecho implica una semejanza de los resultados obtenidos por los expertos y la herramienta desde la perspectiva numérica.

Respecto a las métricas de bondad, la medida Precision (Precisión) resultó ser de 0.82382134, mientras que la métrica Recall (Exhaustividad) alcanzó un valor de 0.741071429 y la ponderada armónica F1 se situó en 0.780258519.

Discusión

Atendiendo a sus resultados, SemSEDoc se presenta como una herramienta que puede producir resultados atractivos para la gestión de repositorios de documentación relativa al proceso de desarrollo de software. Los resultados de la anotación se consideran semejantes a los obtenidos por otras herramientas de anotación basadas en procedimientos de PLN (e.g., García-Crespo *et al.* 2010a; Miao *et al.* 2009), donde la métrica combinada F1 ronda los 0.9 puntos y en algunos casos de forma ligeramente superior a los resultados ofrecidos por otros trabajos del área (Morales del Castillo *et al.* 2009). Con relación a la búsqueda, teniendo en cuenta que ésta basa su efectividad en el anotado de los documentos (se considera madura la tecnología de búsqueda semántica que sustenta la herramienta) los resultados son análogos, aunque ligeramente inferiores a los obtenidos en la anotación. Esto puede ser debido a que, si bien la búsqueda se ha llevado a cabo en un entorno compuesto únicamente por dos documentos (Estudio de viabilidad del sistema y Análisis del sistema de información), la complejidad de ambos documentos es mucho mayor que el documento de Acta utilizado para la primera tarea. Adicionalmente, se detectó que en los proyectos 5 y 8 las diferencias entre las poblaciones provenían de una incompleta definición de los mecanismos de PLN. Dicha incorrección (no inclusión de un sinónimo en el vocabulario), impidió la detección de un total de 6 ocurrencias para el Proyecto 5 y de 7 para el Proyecto 8. Esta incidencia revela un ámbito de mejora sustancial en la gestión de los vocabularios para el PLN. Por otra parte, se considera que los mecanismos semánticos que sustentan el proceso de búsqueda han producido resultados muy destacados en efectividad y rendimiento.

La última de las cuestiones que se deben plantear son las limitaciones del estudio emprendido. La primera de las limitaciones tiene que ver con el ámbito

de la prueba. Dicha prueba se realizó en un entorno de proyectos de tamaño pequeño y controlado. De esta forma, las obligaciones de observancia del rendimiento necesarias para la explotación de SemSEDoc en un entorno real quedan fuera del ámbito del presente trabajo. En segundo lugar, la prueba de búsqueda y anotación también se realizó en un conjunto de documentos definido, por lo que la riqueza de producción documental de un proyecto de desarrollo de software (aún cuando éste se desarrolla guiado por una metodología pesada y perfectamente pautada como Métrica 3) no fue tenida en cuenta en su totalidad. La tercera de las limitaciones proviene de la composición de la muestra, que si bien cubre un total de nueve proyectos, incluye la participación de sólo dos componentes del equipo de trabajo por proyecto.

Conclusiones y trabajos futuros

La importancia del software en el mundo de hoy en día supone que su desarrollo deba ser afrontado desde una perspectiva ingenieril e industrial. En este escenario, la documentación que soporta el proceso software, representa una fuente de información y de gestión del conocimiento que, en muchos casos, no es aprovechada de forma conveniente. Por otra parte, las tecnologías semánticas se han revelado en los últimos años como importantes habilitadoras en aspectos como la gestión del conocimiento y el aprovechamiento inteligente de la información depositada en repositorios. Teniendo en cuenta ambas circunstancias, en el presente trabajo se ha presentado SemSEDoc, una herramienta tendente al aprovechamiento de las ventajas de las tecnologías semánticas en el entorno de la documentación generada como soporte al proceso software.

La herramienta desarrollada permite una gestión de la documentación a partir del anotado semántico semi-automático de los contenidos. Dicho anotado se considera semi-automático, ya que el sistema sugiere las anotaciones, pero es el usuario el que las realiza. Adicionalmente, y desde el punto de vista de búsqueda de informaciones, la aplicación permite la navegación semántica y la búsqueda por facetas de los documentos. Desde el punto de vista de la evaluación de la herramienta, SemSEDoc presenta unos resultados muy prometedores en relación a la gestión de información.

Por último, se quieren poner de manifiesto diversos trabajos futuros que emanan del diseño e implementación de SemSEDoc. En primer lugar, se sugiere la expansión de las funcionalidades del aplicativo para incluir otras metodologías de desarrollo de sistemas de información. En segundo lugar y desde un punto de vista más técnico, se propone la realización de un crawler que permita la gestión de repositorios de artefactos software en formatos distintos a HTML, con una organización no conocida y dispersos desde el punto de vista de su acceso y almacenamiento. En tercer lugar, se propone dotar a la herramienta de capacidades para el análisis de código fuente. Dicha funcionalidad ampliaría las capacidades de utilización de SemSEDoc en entornos donde la documentación no haya sido generada o se haya extraviado, habilitando así el uso de la herramienta en entornos de Reingeniería del Software. Por último, se

13 sugiere la incorporación de tecnologías que permitan el aprendizaje continuado de SemSEDoc en el ámbito de la búsqueda de términos para su posterior anotación. Dichas técnicas relacionadas con la inteligencia artificial permitirían al aplicativo la mejora continuada en el proceso clave para SemSEDoc: las recomendaciones de anotación.

About the authors

Francisco J. García-Peñalvo es Profesor Titular de Universidad en el Departamento de Informática y Automática de la Universidad de Salamanca. Se le puede contactar en el correo electrónico: fgarcia@usal.es

Ricardo Colomo-Palacios es Profesor Titular de Universidad Interino del Departamento de Informática de la Universidad Carlos III de Madrid . Se le puede contactar en el correo electrónico: rcolomo@inf.uc3m.es

Pedro Soto-Acosta es Profesor Titular de Universidad del Departamento de Organización de Empresas y Finanzas de la Universidad de Murcia. Se le puede contactar en el correo electrónico: psoto@um.es

Isabel Martínez-Conesa es Profesora Titular del Departamento de Economía Financiera y Contabilidad de la Universidad de Murcia. Se le puede contactar en el correo electrónico: isabelm.martinez@um.es

Enric Serradell-López es Profesor de los Estudios de Economía y Empresa de la Universidad Oberta de Cataluña. Se le puede contactar en el correo electrónico: eserradell@uoc.edu

- Álvarez Sabucedo, L., Anido-Rifón, L., Corradini, F., Polzonetti, A. & Re, B. (2010). Knowledge-based platform for eGovernment agents: a web-based solution using semantic technologies. *Expert Systems with Applications*, **37**(5), 3647-3656
- Bakry, S.H. & Alfantookh, A. (2010). Toward building the knowledge culture: reviews and a KC-STOPE with six sigma view. *International Journal of Knowledge Society Research*, **1**(1), 47-65
- Benjamins, V.R., Davies, J., Baeza-Yates, R., Mika, P., Zaragoza, H., Greaves, M. *et al.* (2008). Near-term prospects for semantic technologies. *IEEE Intelligent Systems*, **23**(1), 76-88
- Berners-Lee, T., Hall, W., Hendler, J., Shadbolt, N. & Weitzner, J. (2006). Creating a science of the web. *Science*, **313**(5788), 769-77
- Berners-Lee, T., Hendler, J. & Lassila, O. (2001). The semantic web. *Scientific American*, **284**(5), 34-43
- Brooks, T.A. (2009). Watch this: probe the semantic web with SPARQL. *Information Research*, **14**(4) paper TB0912 Retrieved 25 May, 2010 from http://informationr.net/ir/14-4/TB0912.html

Card, D., McGarry, F. & Page, G. (1987). Evaluating software engineering technologies. *IEEE Transactions on Software Engineering*, **13**(7), 845-851

- Cleverdon, C.W., Mills, J., & Keen, E.M. (1966). Factors determining the performance of indexing systems, Cranfield, UK: College of Aeronautics.
- Colomo-Palacios, R., García-Crespo, A., Gómez-Berbís, J.M., Casado-Lumbreras, C. & Soto-Acosta, P. (2010). SemCASS: technical competence assessment within software development teams enabled by semantics. *International Journal of Social and Humanistic Computing*, **1**(3), 232-245
- Colomo-Palacios, R., Gómez-Berbís, J.M., García-Crespo, A. & Puebla-Sánchez, I. (2008). Social global repository: using semantics and social web in software projects. *International Journal of Knowledge and Learning*, **4**(5), 452-464
- Davies, J., Lytras, M.D. & Sheth, A.P. (2007). Semantic-web-based knowledge management. *IEEE Internet Computing*, **11**(5), 14-16
- De Lucia, A., Fasano, F., Oliveto, R. & Tortora, G. (2007). Recovering traceability links in software artifact management systems using information retrieval methods. *ACM Transactions on Software Engineering and Methodology*, **16**(4), 13(article number)
- Dietrich, J., Jones, N. & Wright, J. (2008). Using social networking and semantic web technology in software engineering – use cases, patterns, and a case study. *The Journal of Systems and Software*, **81**(12), 2183-2193
- Frakes, W.B. & Kang, K. (2005). Software reuse research: status and future. *IEEE Transactions on Software Engineering*, **31**(7), 529-536
- Gall, C.S., Lukins, S., Etzkorn, L., Gholston, S., Farrington, P., Utley, D. *et al.* (2008). Semantic software metrics computed from natural language design specifications. *IET Software*, **2**(1), 17-26
- García, F., Ruiz, F., Calero, C., Bertoa, M.F., Vallecillo, A., Mora, B. *et al.* (2009). Effective use of ontologies in software measurement. *The Knowledge Engineering Review*, **24**(1), 23-40
- García-Crespo, A., Colomo-Palacios, R., Gómez-Berbís, J.M. & García-Sánchez, F. (2010b). SOLAR: Social Link Advanced Recommendation System. *Future Generation Computer Systems*, **26**(3), 374-380
- García-Crespo, A., Colomo-Palacios, R., Gómez-Berbís, J.M. & Mencke, M. (2009). BMR: Benchmarking metrics recommender for personnel issues in software development projects. *International Journal of Computational Intelligence Systems*, **2**(3), 257-267.
- García-Crespo, A., Colomo-Palacios, R., Gómez-Berbís, J.M., & Ruiz-Mezcua, B. (2010a). SEMO: a framework for customer social networks analysis based on semantics. *Journal of Information Technology*, **25**(2), 178-188.

- García-Crespo, Á., Gómez-Berbís, J.M., Colomo-Palacios, R. & García-Sánchez, F. (2011). Digital libraries and web 3.0. The callimachusdl approach. *Computers in Human Behavior*, **27**(4), 1424-1430
- Girardi, R. & Leite, A. (2008). A knowledge-based tool for multi-agent domain engineering. *Knowledge-Based Systems*, **21**(7), 604-611
- Gómez-Berbis, J.M., Colomo-Palacios, R., López-Cuadrado, J.L., González-Carrasco, I. & García-Crespo, A. (2011). SEAN: multi-ontology semantic annotation for highly accurate closed domains. *International Journal of the Physical Sciences*, **6**(6), 1440-gon101451
- Gonzálvez-Gallego, N., Soto-Acosta, P., Molina-Castillo, F.J., Trigo, A. & Varajao, J. (2010). El papel de las TIC en el rendimiento de las cadenas de suministro: el caso de las grandes empresas de España y Portugal. *Universia Business Review*, **28**(4), 102-114
- Good, B.M. & Tennis, J.T. (2009). Term based comparison metrics for controlled and uncontrolled indexing languages. *Information Research*, **14**(1), paper 395 Retrieved 25 May, 2010 from http://informationr.net/ir/14-1/paper395.html
- Henriksson, J., Heidenreich, F., Johannes, J., Zschaler, S. & Assmann, U. (2008). Extending grammars and metamodels for reuse: the reuseware approach. *IET Software*, **2**(3), 165-184
- Hernández-López, A., Colomo-Palacios, R., García-Crespo, Á. & Soto-Acosta, P. (2010). Team software process in gsd teams: a study of new work practices and models. *International Journal of Human Capital and Information Technology Professionals*, **1**(3), 32-53
- Hyland-Wood, D., Carrington, D. & Kaplan, S. (2008). Towards a software maintenance methodology using Semantic web techniques and paradigmatic documentation modeling. *IET Software*, **2**(4), 337-347
- Kajko-Mattsson, M. (2005). A survey of documentation practice within corrective maintenance. *Empirical Software Engineering*, **10**(1), 31-55
- Kim, Y. & Stohr, E.A. (1998). Software reuse: survey and research directions. *Journal of Management Information Systems*, **14**(4), 113-147
- Krueger, C.W. (1992). Software reuse. *ACM Computing Surveys*, **24**(2), 131-183.
- Lee, C.S. & Wang, M.H. (2009). Ontology-based computational intelligent multi-agent and its application to CMMI assessment. *Applied Intelligence*, **30**(3), 203-219
- Lethbridge, T.C., Singer, J. & Forward, A. (2003). How software engineers use documentation: the state of the practice. *IEEE Software*, **20**(6), 35-39
- Lopez-Nicolas, C. & Soto-Acosta, P. (2010). Analyzing ICT adoption and use effects on knowledge creation: an empirical investigation in SMEs. *International Journal of Information Management*, **30**(6), 521-

528

- Miao, Q., Li, Q. & Dai, R. (2009). AMAZING: A sentiment mining and retrieval system. *Expert Systems with Applications*, **36**(3), 7192-7198
- Mohagheghi, P. & Conradi, R. (2007). Quality, productivity and economic benefits of software reuse: a review of industrial studies. *Empirical Software Engineering*, **12**(5), 471-516
- Morales-del-Castillo, J.M., Peis, E., Moreno, J.M. & Herrera-Viedma, E. (2009). D-Fussion: a semantic selective disssemination of information service for the research community in digital libraries. *Information Research*, **14**(2) paper 398 Retrieved 25 May, 2010 from http://informationr.net/ir/14-2/paper398.html
- Nicholson, B. & Sahay, S. (2008). Human resource development policy in the context of software exports: case evidence from Costa Rica. *Progress in Development Studies*, **8**(2), 163-76
- O'Sullivan, D. & Dooley, L. (2010). Collaborative innovation for the management of information technology resources. *International Journal of Human Capital and Information Technology Professionals*, **1**(1), 16-30
- Prasad, A.R.D., & Madalli, D.P. (2008). Faceted infrastructure for semantic digital libraries. *Library Review*, **57**(3), 225-234
- Ranganathan, S.R. (1962). *Elements of library classification*. (3rd ed.); New Delhi: Asia Publishing House
- Rombach, H. & Basili, V. (1987). Quantitative assessment of maintenance: an industrial case study. In *Proceedings of the IEEE Conference on Software Maintenance, Austin, Texas, 1987*, (pp. 134-144). New York, NY: IEEE. Retrieved 27 November, 2011 from http://www.cs.umd.edu/users/basili/publications/proceedings/P40.pdf (Archived by WebCite® at http://www.webcitation.org/63VTPeQEq)
- Selby, R.W. (2005). Enabling reuse-based software development of large-scale systems. *IEEE Transactions on Software Engineering*, **31**(6), 495-510
- Shadbolt, N., Hall, W. & Berners-Lee, T. (2006). The semantic web revisited. *IEEE Intelligent Systems*, **21**(3), 96-101
- Sharma, R.S., Ng, E.W.J., Dharmawirya, M. & Samuel, E.M. (2010). A policy framework for developing knowledge societies. *International Journal of Knowledge Society Research,* **1**(1), 23-46
- Soto-Acosta, P., Casado-Lumbreras, C. & Cabezas-Isla, F. (2010a): Shaping human capital in software development teams. The case of mentoring enabled by semantics. *IET Software*, **4**(6), 445-452
- Soto-Acosta, P., Loukis, E., Colomo-Palacios, R. & Lytras, M.D. (2010b): An empirical research of the effect of internet-based innovation on business value. *African Journal of Business Management*, **4**(18), 4096-4105

- Soto-Acosta, P., Martínez-Conesa, I. & Colomo-Palacios, R. (2010c). An empirical analysis of the relationship between IT training sources and IT value. *Information Systems Management*, **27**(3), 274-283
- Soto-Acosta, P. & Meroño-Cerdan, A.L. (2008). Analyzing e-Business value creation from a resource-based perspective. *International Journal of Information Management*, **28**(1), 49-60
- Sugumaran, V. & Storey, V.C. (2003). A semantic-based approach to component retrieval. *ACM SIGMIS Database*, **34**(3), 8-24
- Suominen, O., Hyvönen, E., Viljanen, K. & Hukka, E. (2009). HealthFinland-a national semantic publishing network and portal for health information. *Journal of Web Semantics*, **7**(4), 271-376
- Tappolet, J., Kiefer, C. & Bernstein, A. (2010). Semantic web enabled software analysis. Web Semantics: Science, S*ervices and Agents on the World Wide Web*, **8**(2/3), 225-240.
- Trigo, A., Varajão, J. & Barroso, J. (2009). A practitioner's roadmap to learning the available tools for information system function management. *International Journal of Teaching and Case Studies*, **2**(1), 29-40
- Uren, V.S., Cimiano, P., Iria, J., Handschuh, S.,Vargas-Vera, M., Motta, E. & Ciravegna, F. (2006). Semantic annotation for knowledge management: requirements and a survey of the state of the art. *Journal of Web Semantics*, **4**(1), 14-28.
- Valencia-García, R., García-Sánchez, F., Castellanos-Nieves, D., Fernández-Breis, J. & Toval, A. (2010). Exploitation of Social Semantic Technology for Software Development Team Configuration. *IET Software*, **4**(6), 373-385
- van Rijsbergen, C.J. (1979). *Information retrieval*. Newton, MA: Butterworth-Heinemann.
- Warren, P. (2006). Knowledge management and the semantic web: from scenario to technology. *IEEE Intelligent Systems*, **21**(1), 53-59
- Wongthongtham, P., Chang, E., Dillon, T. & Sommerville, I. (2009). Development of a software engineering ontology for multisite software development. *IEEE Transactions on Knowledge and Data Engineering*, **21**(8), 1205-1217
- Yao, H., Etzkorn, L.H. & Virani, S. (2008). Automated classification and retrieval of reusable software components. *Journal of the American Society for Information Science and Technology*, **59**(4), 613-627
- Zhang, Y., Witte, R., Rilling, J. & Haarslev, V. (2008). Ontological approach for the semantic recovery of traceability links between software artifacts. *IET Software*, **2**(3), 185-203
- Zhao, Y., Dong, J. & Peng, T. (2009). Ontology classification for semantic-web-based software engineering'. *IEEE Transactions on*

Introduction. *Because of the importance of information systems in today's society, their development is a key activity. In this environment, systems are increasingly complex and their development involves the generation of documentation supporting this process, which must be managed effectively and efficiently.*
Method*. Given the need for document repositories to support software development, we propose the use of semantic technologies for cataloguing, labelling, searching and displaying artifacts stored in software repositories.*
Results*. The results of the implementation of these technologies are considered highly satisfactory. On the one hand, the feedback on the techniques used for viewing are positive and, secondly, the tests (precision, recall and F1) are very encouraging for obtaining reliable results.*
Conclusions*. The growing maturity of semantic technologies provide a framework for efficient and effective operation for document repositories generated in software development projects.*